**Credit Risk Analysis Report**

**A trophy with fire coming out of it

Description automatically generated with low confidence**

**Table of Contents:**

1. **Overview of the Analysis**
2. **Model Results**
3. **Summary**

**Overview of the Analysis**

Risk assets issued by lending companies to customers with the anticipation that the borrower perform and pay back as at when due on the principal and the interest elements. Hence, Credit Risk or the risk of default is related to a borrower not paying back or performing on the loan covenants, resulting in lost of income by the lenders.

The risk of a default can be measured by lenders using several methodologies. This case is underpinned by use of Machine Learning to analyze a dataset of historical lending activity from a peer-to-peer lending services company to build a model that can identify the creditworthiness of borrowers.

* Using a machine learning mode 1 & 2, I have been able to assess loans that are healthy (low risk) and non-healthy (high risk) based on the loan status provided by the lending company.
* Under this case, the Logistic Regression Algorithm is the best modeling tool to use for the machine learning model owing to the fact it is widely used to predict the probability of a target variable in classification challenges under ML.

**Based on Machine Learning Model 1**

* Utilizing the whole dataset offered by the financial services company, I have performed a Logistic Regression Modeling with that generated a Balanced Accuracy Score of 95%.
* The classification report for the model are given as 0.85, 0.91,0.88 for precision, recall, F1-score respectively. Even though the model results in a high accuracy, the models recall value (0.91) for non-healthy loans is lower than the recall value (0.99) for healthy loans. The implication is that the model predicts loan status's as healthy better than being able to predict loan status's as non-healthy. This is due to the dataset being imbalanced, meaning that most of the data belongs majorly to one-class label (in this case healthy loans greatly outweighed non-healthy loans).

Graphical user interface

Description automatically generated with medium confidence

* From the code analysis, and the cell titled “Split the data into training and testing datasets by using train\_test\_split”, after coding using the value\_counts function, the print of “y\_train.value\_counts()” reveals that the majority class is healthy loans [0] - 56271 and the minority class is non-healthy loans [1] - 1881:

Graphical user interface, application, Word

Description automatically generated

**Based on the Confusion Matrix**

Text

Description automatically generated

* Out of the 18,719 loan status's that are predicted as healthy (low-risk), the model predicted 18,663 as healthy correctly and 56 as healthy inaccurately.
* Out of the 665 loan status's that are non-healthy (high-risk), the model predicted 563 as non-healthy correctly and 102 as non-healthy erroneously.

**Based on Machine Learning Model 2**

* To predict a higher accuracy score and have the model perform more accurately when classifying non-healthy loans, the original has been oversampled using the RandomOverSampler module from the imbalanced-learn library, which augments the minority class (non-healthy loans) to obtain a balanced dataset.

Graphical user interface, text, application

Description automatically generated

* Further, refitting Logistic Regression Model to the oversampled data, a prediction with an accuracy score of 99%, which turns out to be higher than the model fitted with imbalanced data.
* The oversampled model outperformed because of the dataset being balanced.
* Under this model, the non-healthy loans’ recall value has increased from 0.91 to 0.99, showing that the model has is more robust with an exceptional capacity to detect errors in the labeling of non-healthy (high-risk) loans as healthy (low-risk).

Graphical user interface, application

Description automatically generated

**Based on the confusion matrix**

* Out of the 18,653 loan status's that are healthy, the model predicted 18,649 as healthy correctly and 4 as healthy incorrectly.
* Out of the 713 loan status's that are non-healthy (high-risk), the model predicted 615 as non-healthy correctly and 116 as non-healthy inaccurately.

Text

Description automatically generated

**Results**

* Machine Learning Model 1
* Accuracy score

1. Training Data Score: 0.9921240885954051
2. Testing Data Score: 0.9918489475856377
3. Balanced Accuracy Score: 0.9520479254722232

Graphical user interface, text, application

Description automatically generated

* Precision score

1. precision 1.00 for the healthy loan [0]
2. precision 0.85 for the unhealthy loan [1]

* Recall score

1. Recall 0.99 for the healthy loan [0]
2. Recall 0.91 for the unhealthy loan [1]

Graphical user interface

Description automatically generated with medium confidence

* Machine Learning Model 2:
* Accuracy score

1. Training Data Score: 0.9945659650570917
2. Testing Data Score: 0.9938093272802311
3. Balanced Accuracy Score: 0.9936781215845847

* Precision score

1. precision 1.00 for the healthy loan [0]
2. precision 0.84 for the unhealthy loan [1]

* Recall score

1. Recall 0.99 for the healthy loan [0]
2. Recall 0.99 for the unhealthy loan [1]

Graphical user interface, application

Description automatically generated

**Summary**

* The business model of the lending company is predicated on the assumption that the model categorizes the whole loan portfolio accurate, with health and non-healthy loans labeled properly.
* The failure of the model in the achievement of the objective would bear a grievous monetary consequence on the lending company.
* The Logistic Regression model fitted with OverSampled data performed much better than the model fitted with Imbalanced data due to the data being balanced and generating a higher accuracy score and a higher recall, indicating that the model will make extremely fewer mistakes when classifying non-healthy loans.
* The lending company’s intention would most likely require fewer False Positives due to the high possibility of a lender loosing provided funds when classifying non-healthy loans as healthy.
* Machine Learning Model 1:
* 56 (FALSE POSITIVES): In this case, the actual classification is healthy loans but predicted as non-healthy loan [1].
* 102 (FALSE NEGATIVES): In this case, the actual classification is non-healthy loans, and the predicted as healthy loans [0].
* Machine Learning Model 2:
* 4 (FALSE POSITIVES): In this case, the actual classification is healthy loans but predicted as non-healthy loan [1].
* 116 (FALSE NEGATIVES): In this case, the actual classification is non-healthy loans, and the predicted as healthy loans [0].
* From the confusion matrices, the number of False Positives has drastically decreased, indicating accurate classification by the model of healthy & non-healthy loans.
* Based on the foregoing, I would recommend using Machine Learning Model 2 (Logistic Regression Model fitted with Balanced (oversampled) data.

**References**

Data for this dataset was generated by edX Boot Camps LLC and is intended for educational purposes only.